Extension of a theorem of Boucheron, Lugosi, and Massart

Don Hush and Clint Scovel

Computer Research Group, CIC-3
Los Alamos National Laboratory,
Los Alamos, NM, 87545
(dhush@lanl.gov and jcs@lanl.gov)

August 2, 2000

Concentration of measure has become an important tool in the probabilistic method applied to discrete mathematics, the probabilistic analysis of algorithms, the analysis of randomized algorithms and machine learning. Techniques for proving concentration of measure include the use of Martingale difference inequalities, Talagrand's induction technique, and Marton's use of information theory(See McDiarmid [6] for a survey). However, recently Ledoux [4] has developed a new technique based on logarithmic Sobolev inequalities. In recent work, Boucheron, Lugosi, and Massart [1] used this technique to obtain general concentration of measure results which apply to configuration functions and combinatorial entropies. In this paper we show how the result of Boucheron, Lugosi, and Massart can be extended to determine concentration of measure for the Rademacher statistic and the error deviance, two important functions used in empirical processes and machine learning, (See e.g. Van der Vart and Wellner [7] and Koltchinskii et. al [3, 2]) whose concentration has been obtained through application of Martingale difference inequalities [3, 2].

We begin by stating the theorem of Boucheron, Lugosi, and Massart.

Theorem 1 *(Boucheron, Lugosi, Massart)*

Let $X_1, ..., X_n$ be a set of independent random variables. Let $X = \prod X_j$ denote the product variable and $X^{-i} = \prod_{j \neq i} X_j$ denote the product of all terms except the i-th. Suppose that $Z : X \to R^+$ is a measurable function such that for each i there is some measurable function Z_i of X^{-i} such that

$$0 \leq Z - Z_i \leq 1$$

and

$$\sum (Z - Z_i) \leq Z.$$

Define $h(u) = (1+u)\log(1+u) - u$ for $u \geq -1$. Then for all $\epsilon > 0$,

$$\mathcal{P}(Z \geq E(Z) + \epsilon) \leq e^{-E(Z)h(\frac{\epsilon}{E(Z)})}$$

and for $\epsilon \leq E(Z)$,

$$\mathcal{P}(Z \leq E(Z) - \epsilon) \leq e^{-E(Z)h(-\frac{\epsilon}{E(Z)})}$$

This general theorem can be applied to generate concentration theorems for configuration functions and combinatorial entropies as discussed in [1]. However many random variables satisfy the assumption

$$\sum (Z - Z_i) \leq Z,$$

do not satisfy

$$0 \leq Z - Z_i \leq 1$$

but instead satisfy

$$-\alpha \leq Z - Z_i \leq 1$$

for some $\alpha \geq 0$. We proceed along the lines of Boucheron et al's proof of Theorem 1. The proof is presented in detail for $\alpha = 1$.

Lemma 1 *Let $X_1, ..., X_n$ be a set of independent random variables. Let $X = \prod X_j$ denote the product variable and $X^{-i} = \prod_{j \neq i} X_j$ denote the product of all terms except the i-th. Suppose that $R : X \to R^+$ is a measurable function such that for each i there is some measurable function R_i of X^{-i} such that*

$$|R - R_i| \leq 1$$

and

$$\sum (R - R_i) \leq R.$$

Let

$$F(\lambda) = \log E[e^{\lambda \frac{R}{2} + \frac{n}{2}\phi(-\lambda)}]$$

where $\phi(u) = e^u - 1 - u$ Then

$$F \leq \frac{E(R)}{2}(e^\lambda - 1) + \frac{n}{2}e^{-\lambda}(e^\lambda - 1)^2.$$

Proof. Let $S = \frac{R+1}{2}$ and $S_i = \frac{R_i}{2}$. Then it is easy to see that

$$0 \leq S - S_i \leq 1$$

and

$$\sum (S - S_i) \leq S + \frac{n-1}{2}.$$

S does not satisfy the conditions of Theorem 1. However, Massart [5] shows that

$$\lambda E[Se^{\lambda S}] - E[e^{\lambda S}] \log E[e^{\lambda S}] \leq \sum_i E[e^{\lambda S} \phi(-\lambda(S - S_i))].$$

Since ϕ is convex, $0 \leq S - S_i \leq 1$ and $\phi(0) = 0$

$$\phi(-\lambda(S - S_i)) \leq \phi(-\lambda)(S - S_i)$$

so that

$$\lambda E[Se^{\lambda S}] - E[e^{\lambda S}] \log E[e^{\lambda S}] \leq \phi(-\lambda) E[e^{\lambda S} \sum_i (S - S_i)],$$

but since $\sum (S - S_i) \leq S + \frac{n-1}{2}$ we obtain

$$\lambda E[Se^{\lambda S}] - E[e^{\lambda S}] \log E[e^{\lambda S}] \leq \phi(-\lambda)(E[Se^{\lambda S}] + \frac{n-1}{2} E[e^{\lambda S}]).$$

Rearranging we obtain

$$(\lambda - \phi(-\lambda))E[Se^{\lambda S}] \leq E[e^{\lambda S}](\log E[e^{\lambda S}] + \frac{n-1}{2}\phi(-\lambda)).$$

Partial substitution of $S = \frac{R+1}{2}$ into the left hand side gives

$$(\lambda - \phi(-\lambda))E[\frac{R}{2}e^{\lambda S}] + \frac{(\lambda - \phi(-\lambda))}{2} E[e^{\lambda S}] \leq E[e^{\lambda S}](\log E[e^{\lambda S}] + \frac{n-1}{2}\phi(-\lambda)),$$

which implies that

$$\begin{aligned}(\lambda - \phi(-\lambda))E[\frac{R}{2}e^{\lambda S}] &\leq E[e^{\lambda S}](\log E[e^{\lambda S}] + \frac{n}{2}\phi(-\lambda) - \frac{\lambda}{2})\\ &= E[e^{\lambda S}] \log E[e^{\lambda S + \frac{n}{2}\phi(-\lambda) - \frac{\lambda}{2}}],\end{aligned}$$

which implies that

$$(\lambda - \phi(-\lambda))E[\frac{R}{2}e^{\lambda S + \frac{n}{2}\phi(-\lambda) - \frac{\lambda}{2}}] \leq E[e^{\lambda S + \frac{n}{2}\phi(-\lambda) - \frac{\lambda}{2}}] \log E[e^{\lambda S + \frac{n}{2}\phi(-\lambda) - \frac{\lambda}{2}}],$$

which simplifies to

$$(\lambda - \phi(-\lambda))E[\frac{R}{2}e^{\lambda \frac{R}{2} + \frac{n}{2}\phi(-\lambda)}] \leq E[e^{\lambda \frac{R}{2} + \frac{n}{2}\phi(-\lambda)}] \log E[e^{\lambda \frac{R}{2} + \frac{n}{2}\phi(-\lambda)}] \tag{1}$$

Define

$$F(\lambda) = \log E[e^{\lambda \frac{R}{2} + \frac{n}{2}\phi(-\lambda)}].$$

Then

$$\acute{F} = \frac{E[\frac{R}{2}e^{\lambda \frac{R}{2} + \frac{n}{2}\phi(-\lambda)}]}{E[e^{\lambda \frac{R}{2} + \frac{n}{2}\phi(-\lambda)}]} - \frac{n}{2}\acute{\phi}(-\lambda),$$

so that Equation 1 becomes

$$(\lambda - \phi(-\lambda))(\acute{F} + \frac{n}{2}\acute{\phi}(-\lambda)) \leq F,$$

which amounts to

$$(\lambda - \phi(-\lambda))\acute{F} - F \leq \frac{n}{2}(1 - e^{-\lambda})^2,$$

since $\lambda - \phi(-\lambda) = -\acute{\phi}(-\lambda) = 1 - e^{-\lambda}$.

Observe that $F(0) = 0$ and $\acute{F}(0) = \frac{E[R]}{2}$. Let $v = E[R]$. Then consider $G = \frac{v}{2}(e^{\lambda} - 1)$. Since G satisfies

$$(\lambda - \phi(-\lambda))\acute{G} - G = 0$$

and $G(0) = 0$ and $\acute{G}(0) = \frac{v}{2}$, if we define $\varDelta = F - G$ then $\varDelta$ satisfies

$$(\lambda - \phi(-\lambda))\acute{\mathit{\Delta}} - \mathit{\Delta} \leq \frac{n}{2}(1-e^{-\lambda})^2,$$

with $\mathit{\Delta}(0) = 0$ and $\acute{\mathit{\Delta}}(0) = 0$. From the definition of ϕ this equation can be written

$$(1-e^{-\lambda})\acute{\mathit{\Delta}} - \mathit{\Delta} \leq \frac{n}{2}(1-e^{-\lambda})^2.$$

Further define $\delta = \frac{\mathit{\Delta}}{e^{\lambda}-1}$. It is clear that $\delta(0) = 0$. Then

$$\acute{\delta} = \frac{e^{\lambda}}{(e^{\lambda}-1)^2}((1-e^{-\lambda})\acute{\mathit{\Delta}} - \mathit{\Delta}) \leq \frac{n}{2}e^{-\lambda}.$$

Consequently,

$$\delta(\lambda) - \delta(0) \leq \frac{n}{2}(1-e^{-\lambda})$$

for $\lambda \geq 0$ and

$$\delta(0) - \delta(\lambda) \leq \frac{n}{2}(e^{-\lambda}-1)$$

for $\lambda \leq 0$. Therefore, since $e^{\lambda}-1$ is positive for $\lambda \geq 0$ and negative for $\lambda \leq 0$,

$$\mathit{\Delta} = (e^{\lambda}-1)\delta \leq \frac{n}{2}e^{-\lambda}(e^{\lambda}-1)^2$$

for all λ. Consequently,

$$F = G + \mathit{\Delta} \leq \frac{v}{2}(e^{\lambda}-1) + \frac{n}{2}e^{-\lambda}(e^{\lambda}-1)^2.$$

The proof of lemma 1 is finished.

Theorem 2 *Let $X_1, ..., X_n$ be a set of independent random variables. Let $X = \prod X_j$ denote the product variable and $X^{-i} = \prod_{j \neq i} X_j$ denote the product of all terms except the i-th. Suppose that $R : X \to R^+$ is a measurable function such that for each i there is some measurable function R_i of X^{-i} such that*

$$|R - R_i| \leq 1$$

and

$$\sum (R - R_i) \leq R.$$

Let $h(u) = (1+u)\log(1+u) - u$ for $u \geq -1$. Then for all $\epsilon > 0$,

$$\mathcal{P}(R \geq E(R) + \epsilon) \leq e^{-\frac{E(R)+n}{2}h(\frac{\epsilon}{E(R)+n})}$$

and for $\epsilon \leq E(R) + n$,

$$\mathcal{P}(R \leq E(R) - \epsilon) \leq e^{-\frac{E(R)+n}{2}h(-\frac{\epsilon}{E(R)+n})}$$

Proof. For $\lambda > 0$

$$\mathcal{P}(R \geq E(R) + \epsilon) = \mathcal{P}(e^{\frac{\lambda}{2}(R-E(R))} \geq e^{\frac{\lambda}{2}\epsilon}),$$

which by Markov's inequality gives the bound

$$\mathcal{P}(R \geq E(R) + \epsilon) \leq \frac{E[e^{\frac{\lambda}{2}(R-E(R))}]}{e^{\frac{\lambda}{2}\epsilon}}. \tag{2}$$

In a similar fashion for $\lambda < 0$

$$\mathcal{P}(R \leq E(R) - \epsilon) = \mathcal{P}(e^{\frac{\lambda}{2}(R-E(R))} \geq e^{-\frac{\lambda}{2}\epsilon}),$$

which by Markov's inequality gives the bound

$$\mathcal{P}(R \leq E(R) - \epsilon) \leq \frac{E[e^{\frac{\lambda}{2}(R-E(R))}]}{e^{-\frac{\lambda}{2}\epsilon}}. \tag{3}$$

Consequently, we proceed to bound

$$E[e^{\frac{\lambda}{2}(R-E(R))}].$$

To this end,

$$E[e^{\frac{\lambda}{2}(R-E(R))}] = e^{-\frac{v}{2}\lambda - \frac{n}{2}\phi(-\lambda)} \cdot E[e^{\lambda\frac{R}{2} + \frac{n}{2}\phi(-\lambda)}]$$

which by lemma 1 is bounded by

$$\begin{aligned} &e^{-\frac{v}{2}\lambda - \frac{n}{2}\phi(-\lambda)} \cdot e^{\frac{v}{2}(e^{\lambda}-1) + \frac{n}{2}e^{-\lambda}(e^{\lambda}-1)^2} \\ &= e^{\frac{v}{2}\phi(\lambda) + \frac{n}{2}(e^{-\lambda}(e^{\lambda}-1)^2 - \phi(-\lambda))} \\ &= e^{\frac{1}{2}(v\phi(\lambda) + n\psi(\lambda))} \end{aligned}$$

where $\psi(\lambda) = e^{-\lambda}(e^{\lambda} - 1)^2 - \phi(-\lambda)$. Amusingly,

$$\begin{aligned} \psi(\lambda) = e^{-\lambda}(e^{\lambda} - 1)^2 - \phi(-\lambda) &= e^{-\lambda}(e^{2\lambda} - 2e^{\lambda} + 1) - (e^{-\lambda} - 1 + \lambda) \\ &= e^{\lambda} - 2 + e^{-\lambda} - e^{-\lambda} + 1 - \lambda = e^{\lambda} - 1 - \lambda = \phi(\lambda) \end{aligned}$$

so that

$$E[e^{\frac{\lambda}{2}(R-E(R))}] \leq e^{\frac{1}{2}(v+n)\phi(\lambda)}.$$

Consequently, from the bounds 2 and 3, we obtain that

$$\mathcal{P}(R \geq E(R) + \epsilon) \leq e^{-\frac{1}{2}(\lambda\epsilon - (v+n)\phi(\lambda))}$$

for all $\lambda > 0$ and

$$\mathcal{P}(R \leq E(R) - \epsilon) \leq e^{-\frac{1}{2}(-\lambda\epsilon - (v+n)\phi(\lambda))}$$

for all $\lambda < 0$. The proof of the theorem is finished by showing that

$$\sup_{\lambda>0}(\lambda\epsilon - \xi\phi(\lambda)) = \xi h(\frac{\epsilon}{\xi})$$

for $\epsilon > 0$ and

$$\sup_{\lambda<0}(-\lambda\epsilon - \xi\phi(\lambda)) = \xi h(-\frac{\epsilon}{\xi})$$

for $0 < \epsilon \leq \xi$.

$\square$

The following theorem can be proved in a similar manner to Theorem 2 and provides a continuous interpolation between Theorem 1 and Theorem 2.

Theorem 3 *Let $X_1, ..., X_n$ be a set of independent random variables. Let $X = \prod X_j$ denote the product variable and $X^{-i} = \prod_{j \neq i} X_j$ denote the product of all terms except the i-th. Suppose that $R : X \to R^+$ is a measurable function such that for each i there is some measurable function R_i of X^{-i} such that*

$$-\alpha \leq R - R_i \leq 1$$

and

$$\sum (R - R_i) \leq R$$

with $\alpha \geq 0$. Let $h(u) = (1+u)\log(1+u) - u$ for $u \geq -1$. Then for all $\epsilon > 0$,

$$\mathcal{P}(R \geq E(R) + \epsilon) \leq e^{-\frac{E(R)+n\alpha}{1+\alpha} h(\frac{\epsilon}{E(R)+n\alpha})}$$

and for $\epsilon \leq E(R) + n\alpha$,

$$\mathcal{P}(R \leq E(R) - \epsilon) \leq e^{-\frac{E(R)+n\alpha}{1+\alpha} h(-\frac{\epsilon}{E(R)+n\alpha})}$$

Applications

Here we describe some functions which satisfy the assumptions of Theorems 2 and 3.

Let $R = |\sum \sigma_i \delta_{x_i}|_{\mathcal{F}}$, denote the unnormalized Rademacher statistic where $|h|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |h(f)|$ with $\mathcal{F}$ a class of functions and $\sigma_i, i = 1, .., n$ a vector of Bernoulli random variables taking values 1 and -1 with $p(1) = p(-1) = \frac{1}{2}$. Denote $r = \sum \sigma_i \delta_{x_i}$ so that $R = |r|_{\mathcal{F}}$. Define $r_i = \sum_{i \neq j} \sigma_j \delta_{x_j}$ to be the sum without the i-th term. Then since

$$r = r_i + \sigma_i \delta_{x_i}$$

it is clear that

$$|R - R_i| \leq 1$$

by the convexity of the $\sup_{\mathcal{F}}$ operation. The second assumption follows since

$$r = \frac{1}{n-1} \sum r_i$$

implies that

$$R \leq \frac{1}{n-1} \sum R_i.$$

Consequently, the theorem shows that

$$\mathcal{P}(R \geq E(R) + \epsilon) \leq e^{-\frac{E(R)+n}{2} h(\frac{\epsilon}{E(R)+n})},$$

and for $\epsilon \leq E(R) + n$,

$$\mathcal{P}(E(R) \geq R + \epsilon) \leq e^{-\frac{E(R)+n}{2} h(-\frac{\epsilon}{E(R)+n})}.$$

Divide R by n to obtain the Rademacher statistic and call this new variable R also. Then for the Rademacher statistic R

$$\mathcal{P}(R \geq E(R) + \epsilon) \leq e^{-\frac{n}{2}(E(R)+1) h(\frac{\epsilon}{E(R)+1})}$$

and for $\epsilon \leq E(R) + 1$,

$$\mathcal{P}(R \leq E(R) - \epsilon) \leq e^{-\frac{n}{2}(E(R)+1) h(-\frac{\epsilon}{E(R)+1})}$$

To make comparison with McDiarmid's [6] theorem we use the fact that for $t \geq 0$

$$h(t) \geq \frac{t^2}{2+t}$$

and for $0 \leq t \leq 1$

$$h(-t) \geq \frac{1}{2} t^2.$$

These bounds are obtained by applying the bound

$$\log x \geq 2 \frac{x-1}{x+1}$$

valid for $x = t + 1 \geq 1$ and observing that for the function $\psi(t) = h(-t)$, $\psi(0) = \acute{\psi}(0) = 0$, and $\frac{d^2\psi}{dt^2}(t) \geq 1$ for $0 \leq t \leq 1$. Now,

$$\mathcal{P}(R \geq E(R) + \epsilon) \leq e^{-\frac{n}{2}(E(R)+1) h(\frac{\epsilon}{E(R)+1})}$$

$$\leq e^{-\frac{n}{2} \frac{\epsilon^2}{2(E(R)+1)+\epsilon}},$$

for $\epsilon \geq 0$ and

$$\mathcal{P}(E(R) \geq R + \epsilon) \leq e^{-\frac{n}{2}(E(R)+1) h(-\frac{\epsilon}{E(R)+1})}$$

$$\leq e^{-\frac{n}{4} \frac{\epsilon^2}{E(R)+1}}$$

for $0 \leq \epsilon \leq E(R) + 1$. In both cases the right hand side is always greater than

$$e^{-\frac{n}{4}\epsilon^2}$$

which compares unfavorably by a factor of 2 in the exponent to the estimate

$$e^{-\frac{n}{2}\epsilon^2}$$

obtained through McDiarmid's theorem.

For another application consider the signed error deviance defined as follows.

Let $r(f) = ne(f) - \sum \delta_{x_i}(f)$ and $R = \sup_{f \in \mathcal{F}} r(f)$, where $\mathcal{F}$ is a class of functions and $e(f) = \int f$ is the integral with respect to some unkown probability distribution. r is the signed difference between the integral $e(f)$ and its Monte-Carlo approximation $\sum \delta_{x_i}(f)$, and the signed deviance R is the maximum of this difference over the function class $\mathcal{F}$,. It is important in Machine Learning to have bounds on R. In a similar manner to the Rademacher statistic, we define $r_i(f) = (n-1)e(f) - \sum_{j \neq i} \delta_{x_j}(f)$ and $R_i = \sup_{f \in \mathcal{F}} r_i(f)$. Since $r(f) = r_i(f) + e(f) - \delta_{x_j}(f)$ it is clear that

$$R \leq R_i + \sup_{f \in \mathcal{F}} (e(f) - \delta_{x_i}(f)) \leq 1$$

and

$$R_i \leq R + \sup_{f \in \mathcal{F}} (\delta_{x_i}(f) - e(f)) \leq 1 - e^*$$

where the approximation error

$$e* = \inf_{f \in \mathcal{F}} e(f)$$

is the best error rate one can obtain with the class of functions $\mathcal{F}$. If we let $\alpha = 1 - e^*$, and note that

$$r = \frac{1}{n-1} \sum r_i,$$

R and R_i satisfy the assumptions of Theorem 3

$$-\alpha \leq R - R_i \leq 1$$

and

$$\sum (R - R_i) \leq R.$$

and so the concentration bounds apply. For the error deviance, R is instead defined as $R = \sup_{f \in \mathcal{F}} |r(f)|$. Similar definitions for R_i show that the conditions of Theorem 2 are satisfied.

Acknowledgements. Clint Scovel gratefully acknowledges support from The Los Alamos DOE Program in Applied Mathematical Sciences.

References

1. Boucheron, S., Lugosi, G., and Massart, P., A sharp inequality with applications, *Random Structures and Algorithms* **16**(2000), 277–292.
2. Koltchinskii, V. I., Rademacher Penalties and Structural Risk Minimization, preprint, 1999.
3. Koltchinskii, V. I., Abdallah, C. T., Ariola, M., Dorato, P., and D. Panchenko, Statistical Learning Control of Uncertain Systems: It is Better Than It Seems, *UNM Technical Report* **EECE99-001**(1999)
4. Ledoux, M., On Talagrand's deviation inequalities for product measures, *Probability and Statistics* **1**(1996), 63–87.
5. Massart, P., About the constants in Talagrand's concentration inequalities for empirical processes, to appear in *Annals of Probability*(2000).
6. McDiarmid, C.,, Concentration, *Probabilistic Methods for Algorithmic Discrete Mathematics* Habib, M., McDiarmid, C., Ramirez-Alfonsin, J., Reed, B., Eds., pp. 195–248, Springer-Verlag, Berlin, 1998
7. van der Vaart, A. W., and J. A. Wellner, *Weak Convergence and Empirical Processes*, Springer-Verlag, New York, 1996.

This article was processed using the LaTeX macro package with JNS style